~~Weather, Climatology & Earth Sciences~~

It starts and ends with data

Towards exascale from an earth system science perspective

(Longer versions of this are available in talks on my website
http://home.badc.rl.ac.uk/lawrence

Bryan N Lawrence

University of Reading

NERC SCIENCE OF THE ENVIRONMENT

Science & Technology Facilities Council

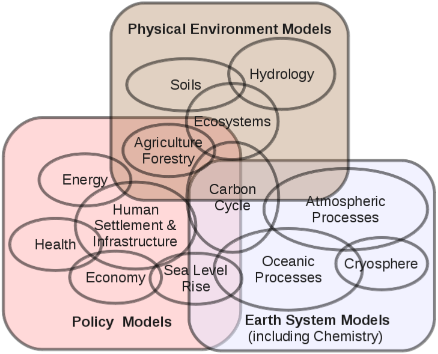
enes European Network for Earth System Modelling

National Centre for Atmospheric Science
NATURAL ENVIRONMENT RESEARCH COUNCIL

Outline

- The Big Picture: Communities and Infrastructure
- Background Trends: Output Data Growth
- Hardware Issues: Storage and Bandwidth
- Software Issues: Analysis software in an exascale world
- Workflow: Bringing compute to the data at scale
- Summary

Communities



Many interacting communities, each with their own software, compute environments etc.

Figure adapted from Moss et al, 2010

Direct Numerical Simulation

Primarily mathematical representation of a complex system of processes

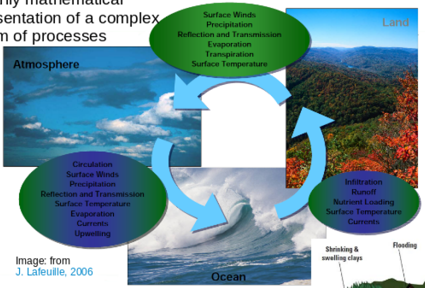
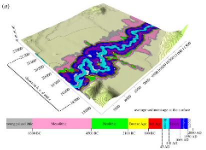
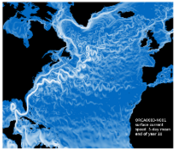
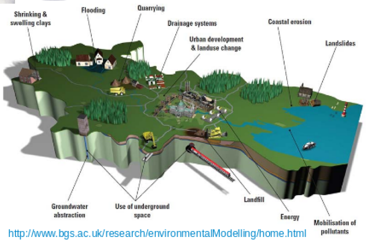


Image: from J. Lafeuille, 2006



Coulthard and Van De Wiel lDoi: 10.1098/rsta.2011.0597



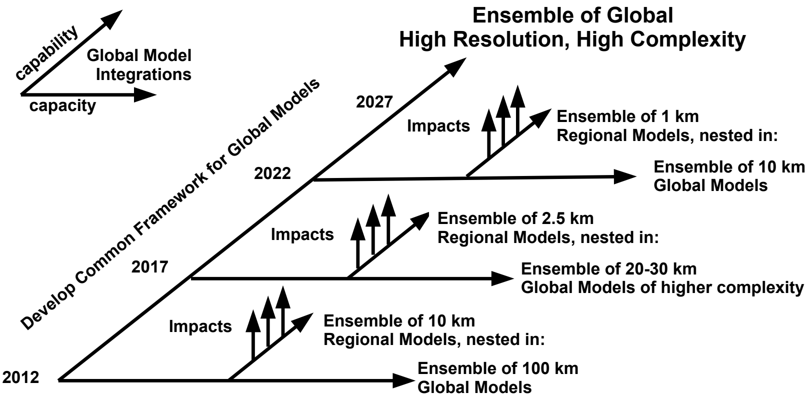


http://www.bgs.ac.uk/research/environmentalModelling/home.html

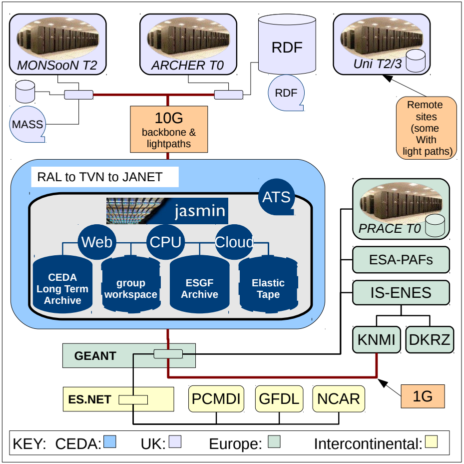
We want to observe and simulate the world at ever higher resolution! More complexity!

Where is this going

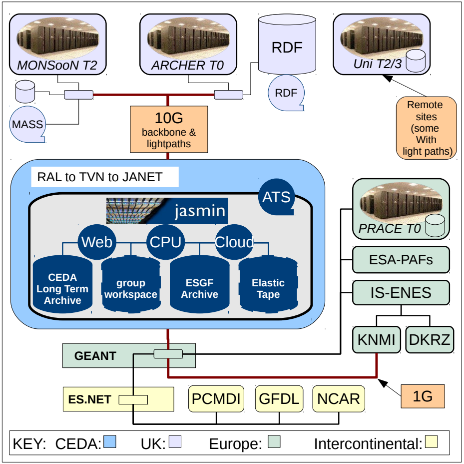
One of many views:



Infrastructure

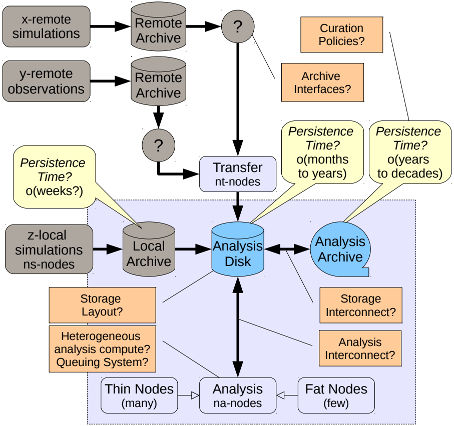


Infrastructure



- The network view is the easy view!
- What are the data policies? What are the (possible) data residence times?
- What agreements are in place?
- What can we rely on in this picture? For example, who has to agree to upgrade something (a network link for example)?
- How do **community** science drivers/requirements lead to infrastructure provision.

An abstract view



- (Potentially) many different remote simulation sources. How long can the data remain at source?
- Interesting problems moving the data to a common location?
- How long can the data reside on disk at the analysis location? What about in the archive?
- How should we best organise the data?
- What are the best ways to organise analysis compute?
- What are the best ways to address analysis interconnect and I/O bandwidth?

Sharing

Science across scales

Lots of interacting communities

Lots of infrastructure

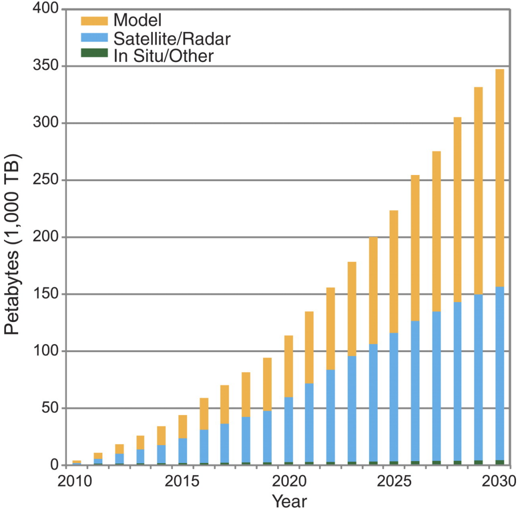
New sorts of infrastructure

Can we share infrastructure?
At exascale, towards exascale?
Between communities?
Between nations?

Global Data Archival

Fig. 2 The volume of worldwide climate data is expanding rapidly, creating challenges for both physical archiving and sharing, as well as for ease of access and finding what's needed, particularly if you're not a climate scientist.

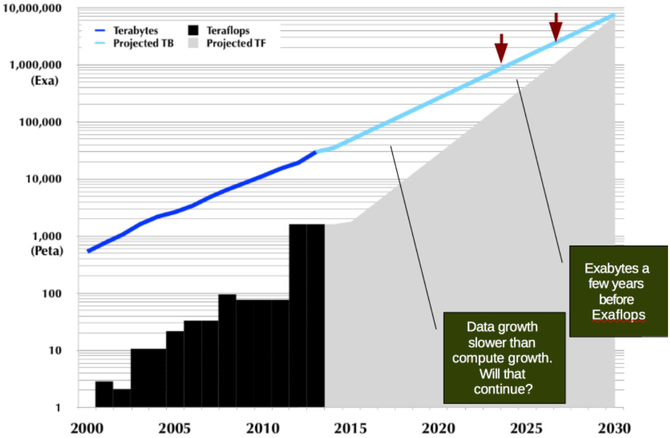
(BNL: Even if you are?)



J T Overpeck et al. Science 2011;331:700-702

Institutional - NCAR

Storage, and power for storage, will dominate NCAR's compute budget within a few years! (Rich Loft, 2014).

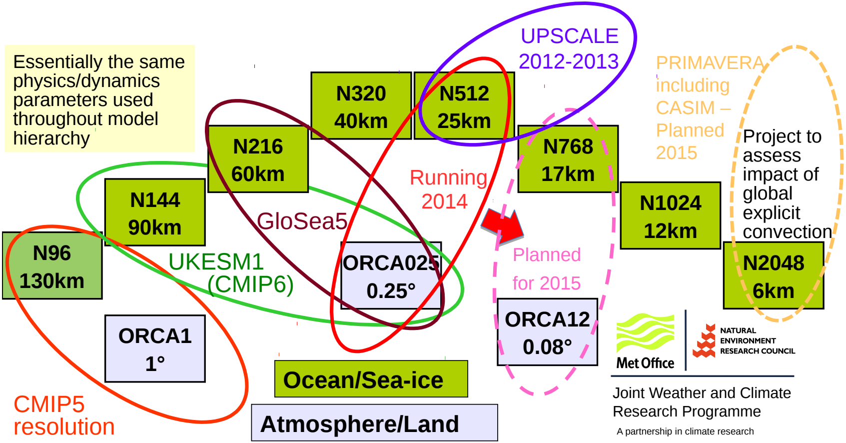


NCAR Compute and Data (courtesy Gary Strand)

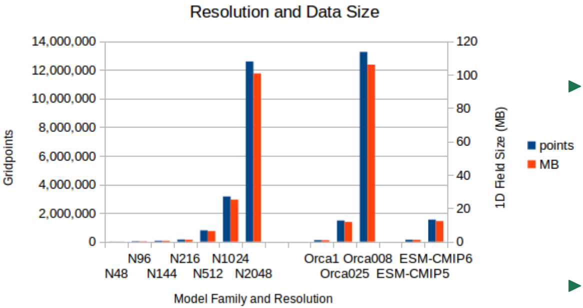
JWCRP Climate Modelling

Earth System Modelling
PI C. Jones (NCAS at the Met Office)

High Resolution Climate Modelling
Joint PIs: P-L. Vidale (NCAS), M. Roberts (Met Office)



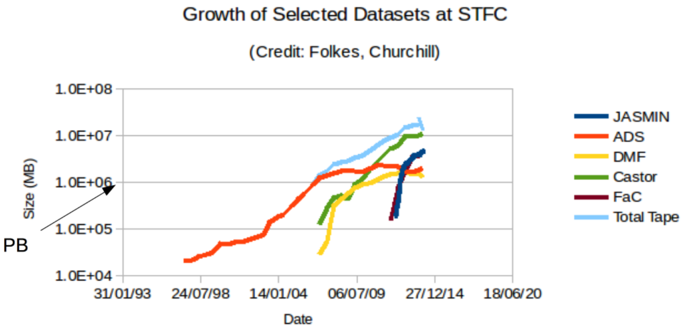
Resolution and Data!



Consequences:

- 1 MB output per 2D field with 10 ensemble members and 100 output variables and 100 levels for $100 * 12 \approx 1000$ time steps $= 10^8$ MB $=$ 100 TB!
- If the UK runs the same number of years for CMIP6 as CMIP5, looks like about o(10) times more data for CMIP6, but could be much worse — more "physics" experiments, means more high resolution experiments, and likely to use bigger ensembles.
- My own experience? Running high resolution gravity wave experiments, 2 years of N512L180 writing data hourly $\approx$ 100 TB. Now!

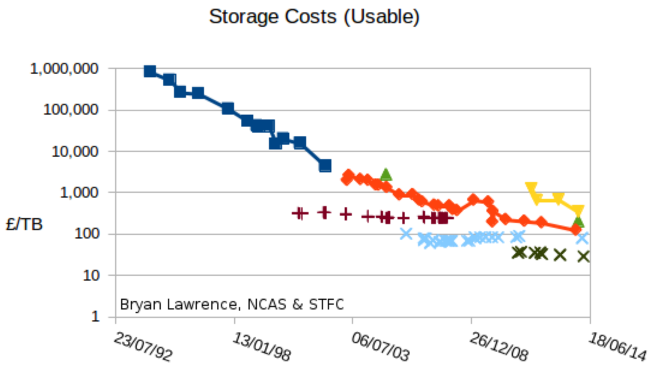
Institutional - STFC and CEDA



Predictions for JASMIN in 2020? 30 — 85 PB of unique data[1]!
But we think we could only fit only 30 PB disk in the physical space available[2]!

([1] Not including CMIP6, which might be anything from 30 PB up. [2] Unless we can throw out the CERC Tier1 centre with whom we share!)

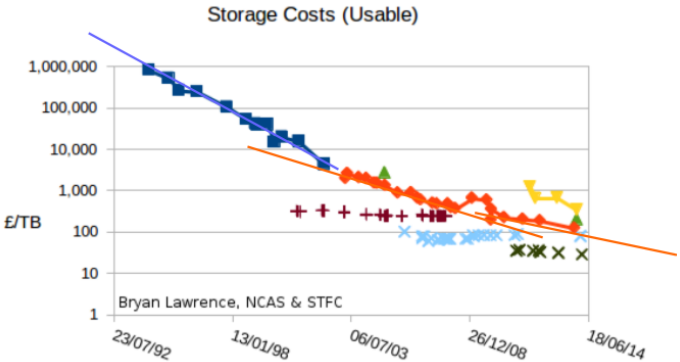
Kryder's Law



Solid objects: colours are different generations of disk. Crosses: different generations of tape.

(Data from Peter Chiu, Jonathan Churchill and Tim Folkes, STFC)

Kryder's Law



Solid objects: colours are different generations of disk. Crosses: different generations of tape.

Kryder's Law definitely slowing down! Plenty of mileage still in tape though!

Storage Density

- Disk: It's getting harder and harder to increase the density of bits on platters, and harder and harder to squeeze platters together.
- Flash: Is competitive, but it seems there is not enough foundry capacity for Flash to take over from disk.
- There will need to be disruptive change to "disk" technology, otherwise physical size of storage will be a problem.
- Tape seems to have more mileage ahead in terms of storage density. Can we make better use of tape in our workflow?

Storage Density

- Disk: It's getting harder and harder to increase the density of bits on platters, and harder and harder to squeeze platters together.
- Flash: Is competitive, but it seems there is not enough foundry capacity for Flash to take over from disk.
- There will need to be disruptive change to "disk" technology, otherwise physical size of storage will be a problem.
- Tape seems to have more mileage ahead in terms of storage density. Can we make better use of tape in our workflow?

Bandwidth to Storage?

(Chicken entail time)

- Historically bandwidth to disk doubling time around 2-3 years, looking forward 6-plus years is possible.
- Bandwidth to tape expected to continue to double at under 3 years?
- The rise of FLAPE? Flash and tape? (No disk!)
- Massive software challenges to use effectively.

Storage Density

- Disk: It's getting harder and harder to increase the density of bits on platters, and harder and harder to squeeze platters together.
- Flash: Is competitive, but it seems there is not enough foundry capacity for Flash to take over from disk.
- There will need to be disruptive change to "disk" technology, otherwise physical size of storage will be a problem.
- Tape seems to have more mileage ahead in terms of storage density. Can we make better use of tape in our workflow?

Bandwidth to Storage?

(Chicken entail time)

- Historically bandwidth to disk doubling time around 2-3 years, looking forward 6-plus years is possible.
- Bandwidth to tape expected to continue to double at under 3 years?
- The rise of FLAPE? Flash and tape? (No disk!)
- Massive software challenges to use effectively.

Not sure how many lessons we can learn from the likes of Google, Facebook etc, even though they are already at silly amounts of storage. Very different access patterns? Different granularity of user volumes?

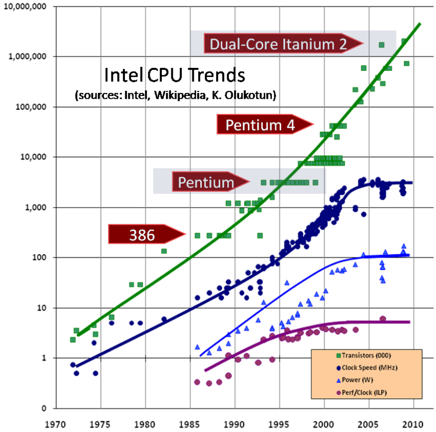
The WAN

Dedicated Lightpath Network

Two weeks in January 2014:
→ Average 10 TB/day, Peak 30 TB/day
→ Inbound onto JASMIN Storage

We've had some network upgrades since then. The bottom line is that we need to, and can, move TBs per day - to JASMIN at least. Looking forward those numbers have to increase. Tens of TB per day in the near future, and PB per day when?

Moore's Law



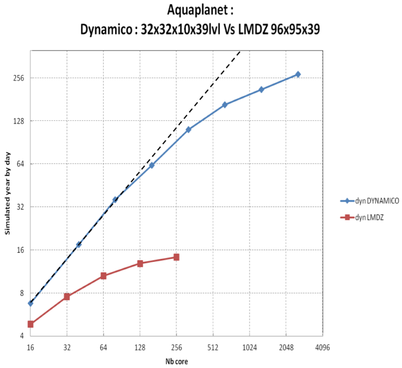
(Herb Sutter, 2004, updated in 2009.)

- Clock speeds not getting any faster (and haven't been for quite a while).
- Transistor density still going up - hence advent of GPUs and accelerators.
- Memory density and bandwidth not keeping up — means it's hard to exploit GPU and accelerators (and going to get harder — fundamental power limits).
- We're kind of used to the problems this means for our simulation codes - massive parallelisation, from MPI to OpenMP to OpenACC ...
- ... problems moving data to exploit the parallelisation etc.

Making progress with parallelisation (1)

Much going on with improving simulation codes

both with coarse parallelisation, for example:



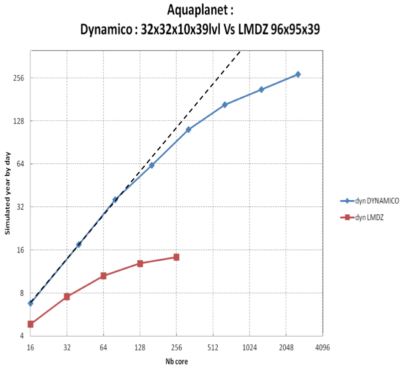
T.Dubos, S.Dubesh, Yann Meurdesoif(LSCE-IPSL)
Results presented at IS-ENES2 workshop, March 2014

is-enes INFRASTRUCTURE FOR THE EUROPEAN NETWORK FOR EARTH SYSTEM MODELLING

Making progress with parallelisation (1)

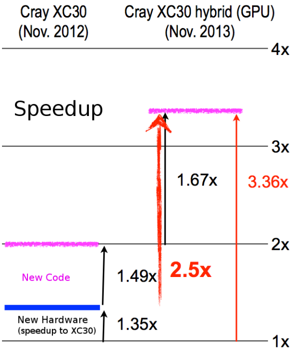
Much going on with improving simulation codes

both with coarse parallelisation, for example:



T.Dubos, S.Dubesh, Yann Meurdesoif(LSCE-IPSL)
Results presented at IS-ENES2 workshop, March 2014

and porting to GPUs, for example:



This work also showed the energy to solution falling by an overall factor of nearly 7 (with a factor of 4 from the GPUs!)

T. Schultess (ETH-Zurich) showing results of 3H Meteo Swiss forecast using the COSMO-2 rewrite, presented at IS-ENES2 workshop, March 2014.

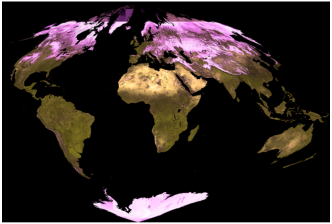
is-enes
INFRASTRUCTURE FOR THE EUROPEAN NETWORK FOR EARTH SYSTEM MODELLING

National Centre for Atmospheric Science
NATURAL ENVIRONMENT RESEARCH COUNCIL

Making progress with parallelisation (2)

Rather less going on with analysis parallelisation

At least much of it is embarrassingly parallel, and we can get results from throwing hardware at the problem, for example:



QA4ECV: "Re-processed MODIS Prior in 3 days (on JASMIN-Lotus). 81 times faster than on 8-core blade".

Boersma and Muller (2014)

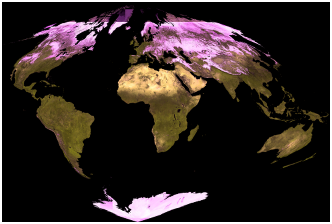
Presentation at http://goo.gl/osEQ6M

From half a year to 3 days!

Making progress with parallelisation (2)

Rather less going on with analysis parallelisation

At least much of it is embarrassingly parallel, and we can get results from throwing hardware at the problem, for example:



QA4ECV: "Re-processed MODIS Prior in 3 days (on JASMIN-Lotus). 81 times faster than on 8-core blade".

Boersma and Muller (2014)

Presentation at http://goo.gl/osEQ6M

From half a year to 3 days!

But we need to work on the software tools (going beyond exploiting queuing or bespoke MPI).
Here for example are some Python choices :

- Standard: Multiprocessing, PyMPI etc
- The way of the future: ipython-notebook
- Generic Workflow and Map Reduce: Jug
- Extending Numpy:
  - Using more cores: Numexpr
  - Using more processors: DistArray (Enthought), Blaze (Continuum Analytics)
- Atmospheric Science aware:
  - PyReshaper, PyAverager (Mickelson, NCAR)
  - cf-python (Hassel, NCAS) (Exploiting LAMA, extending to MPI under the hood soon.)

(Original list courtesy of Matt Jones, UoR)

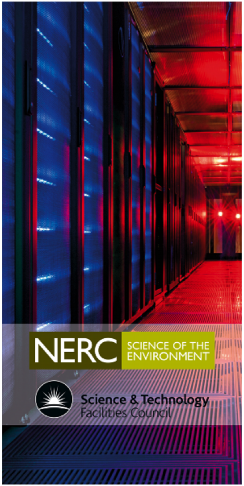
U.S. National Academy

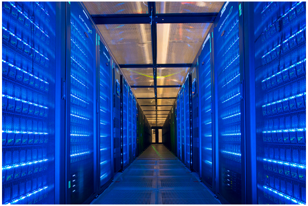
"Without substantial research effort into new methods of storage, data dissemination, data semantics, and visualization, all aimed at bringing analysis and computation to the data, rather than trying to download the data and perform analysis locally, it is likely that the data might become frustratingly inaccessible to users"

A National Strategy for Advancing Climate Modeling, 2012

Semantic Analysis: "substantial research effort" "new methods" "computation to data" "rather than trying to download" "frustratingly inaccessible" (to whom?)

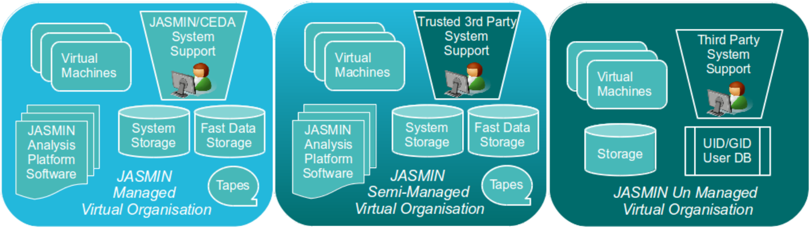
So we have built an "HPC-data" cloud: JASMIN





- 16 PB Fast Storage (Panasas, many Tbit/s bandwidth)
- 1 PB Bulk Storage
- Elastic Tape
- 4000 cores: half deployed as hypervisors, half as the "Lotus" batch cluster.
- Some high memory nodes, a range, bottom heavy.

Virtual Organisations

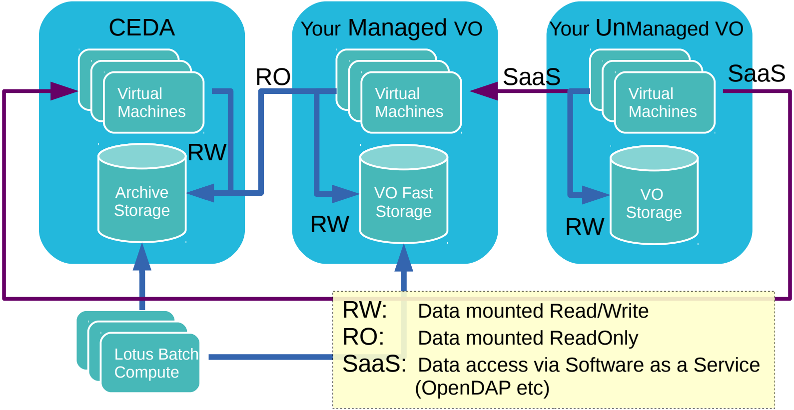


Platform as a Service $\longrightarrow$ Infrastructure as a Service

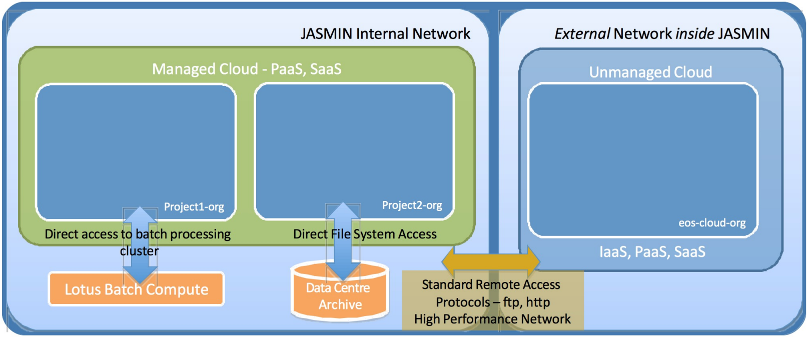
Example: NCAS will run a semi-managed virtual organisation (with multiple group work spaces), but large groups within NCAS can themselves also run virtual organisations.

High performance, curation + facilitation

Objective is to provide an environment with high performance access to curated data archive **and** a high performance data analysis environment!



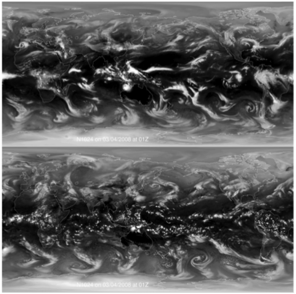
Integrated Cloud Provisioning



Currently o(100) "Group Work Spaces" in the managed cloud serving o(100) "virtual organisations" and o(500) users (there is some overlap). Unmanaged cloud is currently in testing with a few brave souls.

UPSCALE

(Vidale/Roberts - NCAS/Met-Office)

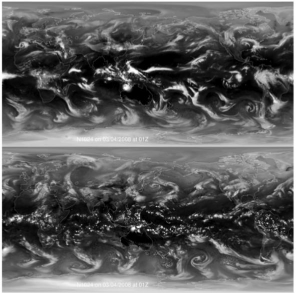


UPSCALE: **U**K on **P**RACE — weather resolving **S**imulations of **C**limate for glob**AL** **E**nvironmental risk.

- Goal: Ensembles of global atmospheric climate simulations at weather forecasting resolution.

UPSCALE

(Vidale/Roberts - NCAS/Met-Office)

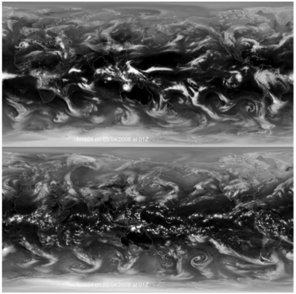


UPSCALE: **U**K on **P**RACE — weather resolving **S**imulations of **C**limate for glob**AL E**nvironmental risk.

- Goal: Ensembles of global atmospheric climate simulations at weather forecasting resolution.
- HPC: Used a one-year 144 million core-hour PRACE allocation on HERMIT (1 PFlop Cray XE6, typically running with up to 50K/115K cores).
- Data: Produced more than 400 TB of data over 10 months, shipped to JASMIN. Expected residence time of core dataset on disk: 5 years.

UPSCALE

(Vidale/Roberts - NCAS/Met-Office)



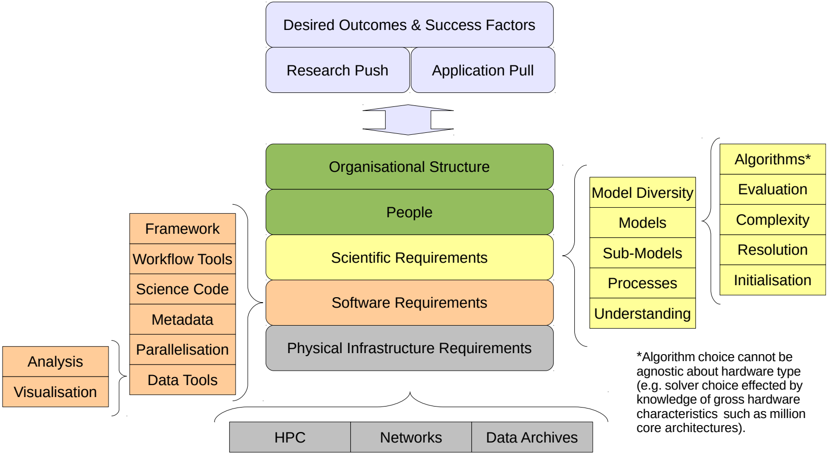
UPSCALE: **U**K on **P**RACE — weather resolving **S**imulations of **C**limate for glob**AL** **E**nvironmental risk.

- Goal: Ensembles of global atmospheric climate simulations at weather forecasting resolution.
- HPC: Used a one-year 144 million core-hour PRACE allocation on HERMIT (1 PFlop Cray XE6, typically running with up to 50K/115K cores).
- Data: Produced more than 400 TB of data over 10 months, shipped to JASMIN. Expected residence time of core dataset on disk: 5 years.
- Access: UPSCALE data **i**nitially accessed via two VMs: one managed by the met office, one by NERC, with 25 & 33 users respectively — a total of 50 unique data users (11/2014).
- HPC: Data analysis on the Lotus cluster - thousands of data analysis cores, PBs of fast disk.

Linpack is nearly useless!

- Many important codes cannot exploit accelerators (either sort).
- Much more important (for us) to understand "SYPD": Simulated Years Per (real) Day — for a given code — from when you typed run to when the last history file was archived.
  - THEN, in the context of BDEC, you need to understand the analysis workflow, and how it will be supported.

Putting it all together



*Algorithm choice cannot be agnostic about hardware type (e.g. solver choice effected by knowledge of gross hardware characteristics such as million core architectures).

Final Remarks

- When we consider the entire workflow associated with environmental simulation, we realise that the "time in the supercomputer" **doing** simulation, is only a small part of the entire workflow.
- When we look at the trend in the balance of hardware spending at *weather and climate* supercomputing sites we see a trend towards a greater proportion of the funding on the storage, but

Final Remarks

- When we consider the entire workflow associated with environmental simulation, we realise that the “time in the supercomputer” **doing** simulation, is only a small part of the entire workflow.
- When we look at the trend in the balance of hardware spending at *weather and climate* supercomputing sites we see a trend towards a greater proportion of the funding on the storage, but
- We have yet to see a comensurate trend towards the spend for an appropriate software infrastructure for data, and
- We have yet to see a real understanding of the data handling implications at the generic national and international facilities, although they’re all beginning to recognise there might be a problem!

Final Remarks

- When we consider the entire workflow associated with environmental simulation, we realise that the “time in the supercomputer” **doing** simulation, is only a small part of the entire workflow.
- When we look at the trend in the balance of hardware spending at *weather and climate* supercomputing sites we see a trend towards a greater proportion of the funding on the storage, but
- We have yet to see a comensurate trend towards the spend for an appropriate software infrastructure for data, and
- We have yet to see a real understanding of the data handling implications at the generic national and international facilities, although they’re all beginning to recognise there might be a problem!

The bottom line: Getting our models to run on (new) supercomputers is hard. Getting them to run performantly is hard. Analysing, exploiting and archiving the data is (probably) **now** even harder!